FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Assessing the impact of alternative splicing in cancer

Ana Gomes

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rui Camacho (FEUP)

Second Supervisor: Valdemar Máximo (FMUP/IPATIMUP)

July 26, 2015

Assessing the impact of alternative splicing in cancer

Ana Gomes

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: João Moreira
External Examiner: Sérgio Matos
Supervisor: Rui Camacho

July 26, 2015

Abstract

Worldwide, millions of people live every day with a diagnosis of cancer. Cancer has several possible causes. One of such possibilities is the genomic origin. In this thesis we designed and developed informatics tools to help cancer researchers to investigate cancer possible origin an [aberrant] alternative splicing. In this process usually a single fragment of DNA can result in more than one transcript during which an aberrant mutation can occur and be the cause of a disorder.

For the genome analysis RNA-seq was used in our study. RNA-seq has been used nowadays, quite frequently, as a procedure to sequence genomes. RNA-seq performs the reconstruction of at least part of the genome of a patient from small fragments of it (reads), calculates the set of active genes and compares it with one from a reference person. This last step of active gene differentiation may help researchers to understand the original biological question that triggered the study. At this last stage it is also important to collect several kinds of information associated with the active genes in order to establish a solid base for informed decisions based on the process.

Although the tools to achieve this evaluation do exist, usually they are dispersed causing the process to be difficult and slow. The whole process requires considerable computational resources and programming skills. Furthermore, it is important for the scientist to work with a user-friendly web interface.

Our main purpose is to develop an application that helps researchers in this task of assessing the impact of [aberrant] alternative splicing in cancer by automating the full process from the reads analysis up to the results of alternative splicing analysis. To achieve those, the work includes the following tasks: develop a web interface to simplify the analysis process, assemble the existing iRAP pipeline and improving the gene enrichment step. Our contribution is four fold: make the whole process easy to use by the biologist expert; design and deploy the data analysis steps; extend an existing pipeline with module(s) specific for splicing; and apply our work in IPATIMUP's data on cancer. Automatization is the major contribution to improve efficiency and quality of the scientific research on the impact of alternative splicing in cancer.

Resumo

Por todo o mundo, milhões de pessoas vivem diariamente com um diagnóstico de cancro. Um dos possíveis fenómenos que se suspeita estar na origem de alguns cancros é o *alternative splicing*. Este processo ocorre no início da transcrição de DNA em RNA. Durante este processo, normalmente uma pequena região de DNA (um gene) pode resultar em mais de uma sequência alternativa de RNA. Mutações ocorridas na sequência de DNA podem ser nefastas e estar na origem de certos tipos de cancro.

Para a análise de genomas utilizamos a tecnologia RNA-seq. RNA-seq é uma tecnologia cada vez mais utilizada para estudar o problema acima descrito. O RNA-Seq executa a reconstrução de pelo menos parte do genoma de um paciente a partir de pequenos pedaços do mesmo (*reads*), calcula o conjunto dos genes ativos e compara-os com um de um grupo de referência. O último passo do processo é geralmente a diferenciação dos genes ativamente expressos o que pode ajudar os investigadores a perceberem a origem biológica da doença. Nesta última etapa é também importante agregar diversos tipos de informação relacionados com os genes ativos com o objetivo de constituir uma base sólida para sustentar as explicações científicas.

Apesar de as ferramentas utilizadas nesta avaliação estarem disponíveis, normalmente estas encontram-se dispersas pela Web, o que torna o processo lento e de difícil execução. Mesmo computacionalmente é uma metodologia que requer recursos consideráveis e competências de programação. Além disso, é importante para o especialista interagir com uma interface amigável e que lhe permita a visualização dos resultados.

O principal objetivo do trabalho é desenvolver uma aplicação que ajude os investigadores nesta tarefa de avaliar o impacto do *alternative splicing* em cancro através da automatização de todo o processo desde a análise dos *reads* até aos resultados da análise do *alternative splicing*. Para o conseguir alcançar, o plano de trabalhos inclui as seguintes tarefas: desenvolvimento de uma interface web para a simplificação do processo de análise, montagem da *pipeline* iRAP existente e melhoria do passo de *gene enrichment*. A nossa contribuição tem quatro vertentes: simplificar o processo para o investigador; planear e implementar os passos da análise de dados; extender a *pipeline* existente com um módulo específico para o *splicing*; e aplicar o trabalho em dados do IPATIMUP sobre cancro. A automatização é a maior contribuição no sentido de melhorar a eficiência e qualidade da investigação científica no que respeita ao impacto do *alternative splicing* no cancro.

Acknowledgements

Foremost, I would like to express my gratitude to my supervisor Prof. Rui Camacho (FEUP) for the continuous support, patience and motivation. His guidance helped me in all the time of develop, research and writing of this thesis. I also would like to thank Prof. Valdemar Máximo (FMUP/IPATIMUP) and IPATIMUP's staff for collaborating and providing the necessary material for case studies analysis.

I thank my fellow college mates of MIEIC for this 5-year journey of companionship as well as my roommates that stood out for me at every time.

Furthermore, my sincere thanks goes to my family: my parents and my brother for their faith in me and allowing me to be as ambitious as I wanted. Last but not the least, I would like to thank my boyfriend, for his love, encouragement, quiet patience and also his help on surpassing all the obstacles I had.

Ana Gomes

"It always seems impossible,
until it's done."

Nelson Mandela

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

BAM	Binary Sequence Alignment/Map
CSS	Cascading Style Sheets
DNA	DesoxyriboNucleic Acid
ENCODE	Encyclopedia of DNA Elements
GFF	General Feature Format
GTF	General Transfer Format
HTML	HyperText Markup Language
mRNA	Messenger RNA
ORF	Open Reading Frame
RNA	RiboNucleic Acid
SAM	Sequence Alignment/Map
SQL	Structured Query Language
WSGI	Web Server Gateway Interface
mFTC	minimially invasive Follicular Thyroid Carcinoma
wFTC	widely invasive Follicular Thyroid Carcinoma

Chapter 1

Introduction

1.1 Context

Worldwide, millions of people live every day with a diagnosis of cancer. One of several possible causes of this disease is an anomalous modification in the process of gene transcription [CH07], more precisely in the early steps of the transcription process where the genes (DNA) are translated to RNA. It is common for a single gene to be transcribed to more than one transcript. This is designated as alternative splicing [Joh03]. It happens, seldom, that errors may occur and an aberrant transcript can be generated and cause a disorder.

In order to study this scientific problem, researchers are now using high throughput technologies that provide them the ability to perform large scale experiments that otherwise would not be humanly possible. One of such technologies is called RNA-Seq [ZW09] in which its last step of active gene differentiation may help researchers to understand the original biological question that triggered the study.

Although RNA-seq is a very valuable asset for genomic research, it is quite often that further analysis need to be performed to answer the initial research question. Commonly those analysis require the recollection of a lot of information stored in a diverse number of data bases spread in the Internet. This is often a very time consuming task for a biologist expert. Moreover there are some studies for which special tools/programs have to be developed to process the output of RNA-seq to answer specific questions of the study.

1.2 Motivation and Goals

Since cancer is a worldwide disease that kills a large amount of people each year, any developments in the treatment of such disease may have a large and very important social impact. High-throughput technologies like RNA-Seq represent already a major improvement for studies on the genomic origins of cancer. They have however some limitations that prevent them to be used in

large scale by geneticists and Molecular Biologists. They require both: powerful computational resources and programming and operating systems skills. In order to help the biologist expert to formulate useful scientific hypotheses, genetic-based cancer studies require also the collection of a large amount of extra information associated with the expressed genes. This last task is usually done by hand over the Internet and is very time consuming.

Our goals are threefold: to automate, as much as possible, the full process of RNA-seq analysis up to the [aberrant] alternative splicing analysis; to make the whole process easy to perform, including the data collection over the Internet; and to extend the available RNA-seq tools with software for extra special purposes analysis.

1.3 Proposed Solution

Our proposal consists in the design, development and deployment of a computational framework that enables expert biologists to control the whole analysis process over a Web interface. The framework also includes the control and use of powerful computational resources in a transparent way for the user.

We will use the specific domain problem of assessing the impact of alternative splicing in cancer as a use and test case by automating the full process from the reads analysis up to the results of alternative splicing analysis. It is also important to provide information enrichment that might be useful to explain the occurrence of cancer originating from the alternative splicing. To achieve those, the work includes the two main tasks: design, develop and deploy a web interface to simplify the analysis process and assemble the existing iRAP pipeline. Our contribution is four fold: make the whole process easy to use by the biologist expert, design and deploy the data analysis steps, extend an existing pipeline with a module specific for splicing and apply our work in IPATIMUP's data on cancer. Automation is the major contribution to improve efficiency and quality of the scientific research on the impact of alternative splicing in cancer.

1.4 Structure of the Dissertation

Besides this introductory chapter this report has the following structure. Chapter 2 presents a review of the state-of-the-art describing not only some tools from the molecular biology domain but also the main biological base concepts. It is also demonstrated the standard file formats used as well as the relationship between relational and non-relational databases. Lastly there is also a brief overview of the related work.

In Chapter 3 the proposed framework, called Gemini, is described. The framework includes a Web interface, a data base, an ftp server and a computational resources server. The chapter starts with the solution description of Gemini. Afterwards, the details of implementation are described and finalizes with the characterization of the case studies done for validation purposes.

Chapter 4 introduces the case study where we have used real world data provided by IPATIMUP's researchers. It starts with the samples characterization, researching objectives and protocol of analysis. Following the data results are described.

The final conclusions of this Dissertation are then presented in Chapter 5. In this last chapter we also point out further directions to extend the work done in the thesis. The framework is conceived in a way that extra modules may be developed and integrated in order to enable other more specific analysis.

Chapter 2

Basic Concepts and Bibliographic Review

In this chapter we present the biological basic concepts necessary to understand the thesis work and we explain what it is gene expression and a more in-depth explanation of the alternative splicing. This will be followed by a state-of-the-art review of some different tools available to analyse RNA and then a characterization of a pipeline till the gene expression results. Some data repositories are presented together with the data formats that might be used. Finally, we describe and analyse some the important technologies used, and describe the main differences between relational and non-relational databases.

2.1 Biological Basic Concepts

We start introducing some of biological basic concepts from the field of molecular biology.

2.1.1 Genome

The flow of genetic information is known as the central dogma of molecular biology. According to it, RNA molecules are synthesized from DNA templates (transcription) and proteins are synthesized from RNA templates (translation) [CH07].

RNA is a single-stranded molecule that is responsible for synthesizing the proteins of the cell. On the other hand, DNA is a double-stranded molecule that carries the genetic information in all cellular forms of life. It can be copied or 'replicated', as each strand can act as a template for the generation of the complementary strand.

Furthermore, in DNA the information is stored in the linear sequence of the nucleotides along each strand. Each sequence of these nucleotides is called a gene that later on the translation process may specify amino acids, the elements of a protein.

As such, genes consist of three types of sequences:

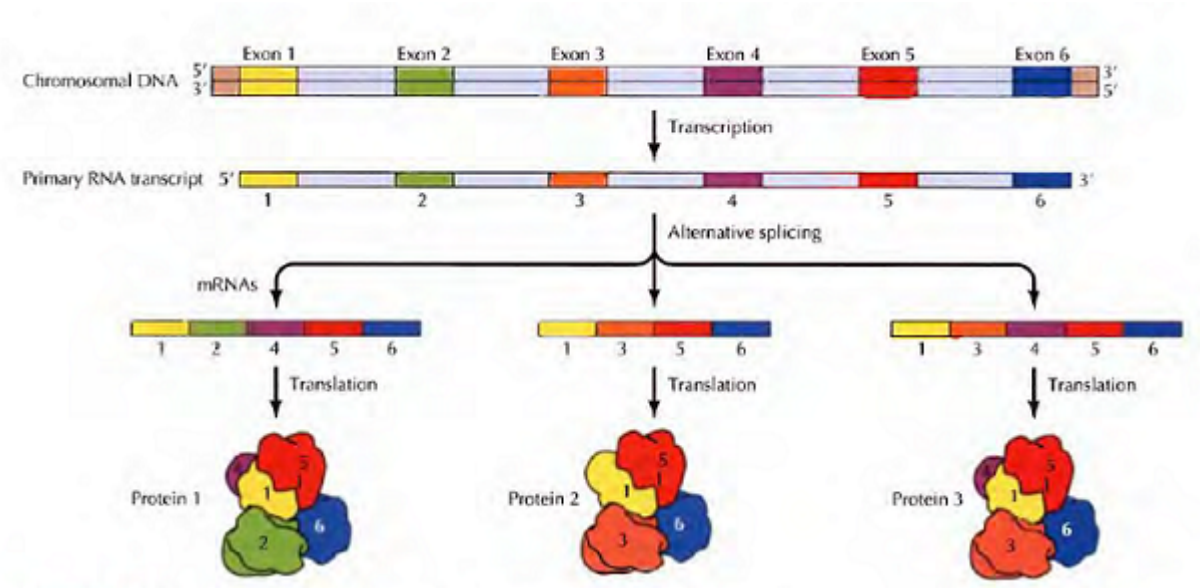

Figure 2.1: Alternative splicing allows exons to be joined in different combinations, resulting in the formation of three distinct proteins from the single primary transcript.

- Coding regions, called exons, which specify a sequence of amino acids;
- Non-coding regions, called introns, which do not specify amino acids;
- Regulatory sequences, which play a role in determining when and where the protein is made.

In the process of splicing, during the transcription, the introns are removed and only exons are included in the messenger RNA that will serve as basis of the production of some protein. Even though these introns do not carry useful information for further steps, they allow the exons of a gene to be joined in different combinations resulting in the synthesis of different proteins from the same gene, as seen in Figure 2.1. This mechanism is called alternative splicing and provides an important tool for tissue-specific control of gene expression in complex cells [CH07].

During this process the nucleotides of DNA are transcribed to RNA ones through a par-based code. Each nucleotide of DNA has on its constitution one of the four bases: Adenine (A), Guanine (G), Thymine (T) or Cytosine (C). On the other hand, a RNA strand is similar but the four possible bases are: Adenine (A), Guanine (G), Uracil (U) and Cytosine (C). The Table 2.1 relates each base of DNA to RNA.

DNA base	**RNA base**
Adenine (A)	Uracil (U)
Thymine (T)	Adenine (A)
Guanine (G)	Cytosine (C)
Cytosine (C)	Guanine (G)

Table 2.1: DNA nucleotides corresponding to RNA ones.

On RNA, its four-letter 'alphabet' forms 'words' of three letters called codons. Individual codons code for specific amino acids. The genetic code is a table that relates the codons to the corresponding amino acids (see Figure 2.2 [1]). One important information that we can retrieve from this table is the start and the stop codons. There is only one start codon, the Methionine (*AUG* combination) and it defines where the translation starts. On the other hand, there are three possible combinations for the stop codon (*TAA*, *TAG* and *TGA*) whose function is to determine where the translation ends [GEN]. Sometimes, due to mutations, the stop codon is not present or misplaced in the strand due to an aberrant alternative splicing. When this happens, it may be created an abnormal protein.

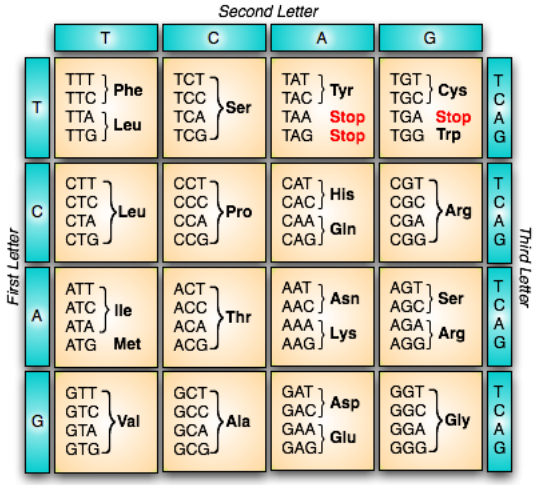

Figure 2.2: Genetic code table.

Another important concept is the Open Reading Frame (ORF). An ORF is a portion of DNA that, when translated into amino acids, has no stop codons. A long open reading frame is likely part of a gene that will serve as basis to encode a protein.

A fusion gene is a hybrid gene that is formed from two previously separated genes. They can result by either gene translocation, interstitial deletion or chromosomal inversion (see Figure 2.3). Nowadays, 358 gene fusions involving 337 different genes have been identified and an increasing number of gene fusions are being recognized as an important diagnostic and prognostic parameters in malignant haematological disorders and childhood sarcomas. An analysis of available data shows that gene fusions occur in all malignancies, and that they account for 20% of human cancer morbidity [SOdMVN12].

2.1.2 Transcriptome Assembly and Gene expression

After the collected samples, the next step is to prepare the library for further analysis. It is common for Biologists to use Illumina [2] for this task.

[1] Image source: http://plato.stanford.edu/entries/information-biological/GeneticCode.png

[2] Illumina: http://www.illumina.com/

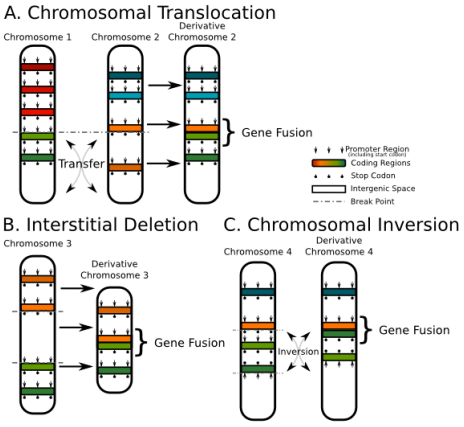

Figure 2.3: Fusion genes generation types.

During the procedure it is performed a process called sequencing [3] that determines the order of nucleotide bases within a stretch of DNA. Next-generation sequencing (NGS) extends this technique across millions of reactions and involves rapid sequencing of large DNA stretches spanning entire genomes.

On particular for RNA, Next-generation RNA sequencing enables researchers to identify and quantify both rare and common transcripts; align sequencing reads across splice junctions, and detect isoforms, novel transcripts and gene fusions; and perform robust whole-transcriptome analysis on a wide range of samples, including low-quality ones [4].

A transcriptome is the a set of all RNA molecules transcribed from a RNA template [Wol13]. From the transcriptome, and according to the used technique, it is possible to count the number of transcripts and then determine the amount of gene activity (gene expression) in a certain cell or tissue type. This is particularly useful when researchers are trying to discover the function of a certain gene: by analysing a transcriptome and the corresponding gene expression levels, it can be inferred what role does it play whether it is in cell growth or in fat storage, for example [Ins]. As such, transcriptome assembly is a computational reconstruction of RNA from smaller sequences, called reads and obtained experimentally.

Differently from the genome, the transcriptome actively changes according to context. In order to follow this changes there is a laboratory technique called microarray that measures the expression of thousands of genes at the same time providing gene expression profiles [bNE].

The assembly can be done by either aligning the previous libraries obtained from Illumina to a reference genome or reference transcripts, or assembled *de novo* without the genomic sequence to produce a genome-scale transcription map that consists of both the transcriptional structure and or level of expression for each gene [ZW09].

[3] As in Illumina techniques description at http://www.illumina.com/applications/sequencing.html
[4] As seen in: http://www.illumina.com/applications/sequencing/rna.html

2.2 RNA-Seq Analysis

As stated in the previous section there are alternative techniques to obtain the gene expression. In this section we compare them and then provide detailed information on RNA-Seq.

2.2.1 RNA-Seq, microarrays and DNA sequencing

For the analysis of genome expression it is commonly used RNA sequencing. Firstly, it is chosen this technique instead of DNA sequencing because some molecular features, like alternative isoforms, can only be observed at the RNA level, predicting transcript sequence from genome is difficult and in functional studies even though genome may be constant, an experimental condition has a pronounced effect on gene expression. Also, it is a way to interpret mutations that do not have an obvious effect in the protein sequence but affect what mRNA isoform is expressed and how much [Gri].

Secondly, even though microarrays have facilitated gene expression based analysis but provide relatively little information about alternative splicing. In late 90's, microarray experiments were generally expensive, limiting sample sizes which represented a disaster specially when associated to thousands of independent observations per sample. Some attempts have been made to surpass this problem but none could perform well in brain tissue. RNA-Seq then appeared as a method to mapping and quantifying transcriptomes. It has clear advantages over existing approaches since it has a greater dynamic range, detects both coding and noncoding RNAs, is superior for gene network construction, detects spliced transcripts and allele specific expression and can be used to extract genotype information. Also by comparing both of them while microarray output is fluorescence intensity, the output from a RNA-Seq experiment is digital and comes in the form of read counts [HBD+13].

Furthermore, RNA-Seq provides a far more precise measurement of levels of transcripts and their isoforms than other methods. It can reveal precise location of transcription boundaries and is particularly useful to apply on complex transcriptomes. In addition, RNA-Seq is shown to be highly accurate for quantifying expression levels while requiring less RNA sample when compared to another techniques [ZW09] and it also can resolve both gene expression level and alternative splicing events simultaneously [GCW+10].

2.2.2 RNA-Seq Pipeline

RNA-Seq has become the tool of choice for genome-wide analysis of the transcriptome [ORY10, SG15]. However, a typical RNA-Seq experiment generates millions of raw sequence reads that require considerable computational resources and programming skills to process the data [GTBK11, ORY10, FPMB14, FMB14]. An overview of a basic pipeline for such analysis can be found in Figure 2.4.

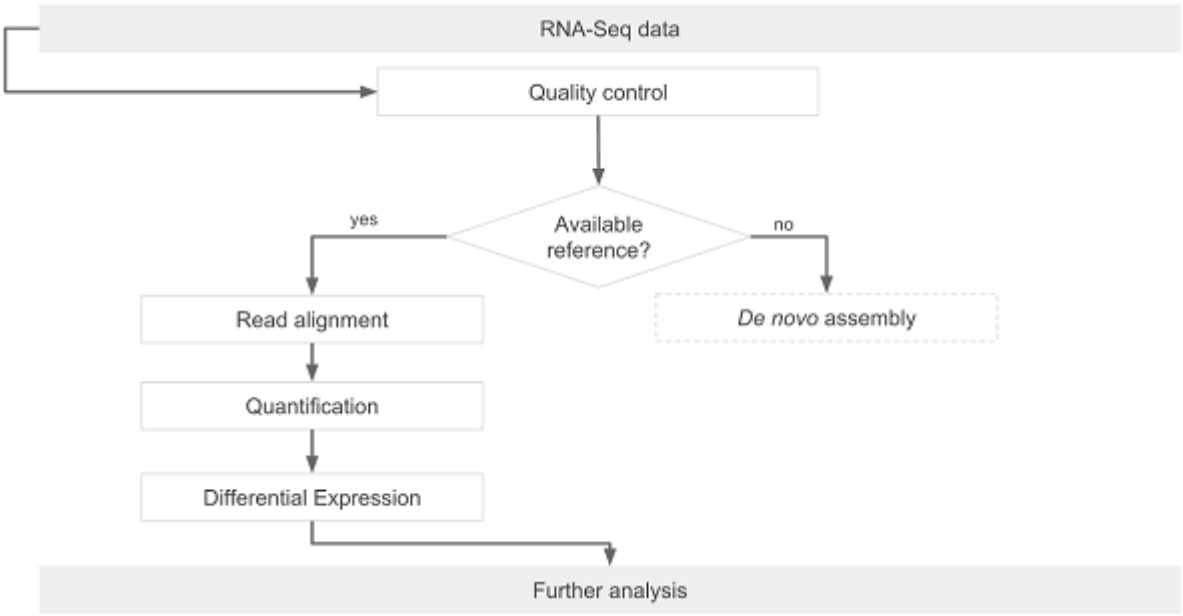

Figure 2.4: The essential workflow for RNA-Seq analysis. [kn:11]

When no reference genome is available it is necessary to construct a *de novo* assembly. This step will not be considered here since it was not used on the thesis work. Given this we are able to hold this main concepts:

Quality control a pre-processing step where quality of raw reads are assessed. At this stage bad "quality" ones may be excluded from the data set preventing potential mapping mismatches and providing a better performance in further states.

Read alignment that aligns the (quality filtered) reads to a reference genome in order to reconstruct the larger sequence.

Quantification that summarises and aggregates reads over a biologically meaningful unit such as exons, isoforms or genes.

Differential expression that identifies genes expressed being the basis for future assessing of proteins function originating from such genes.

Note that this is just a basis workflow for RNA-Seq analysis that provide the possibility to add or remove some of the stages according to different purposes.

iRAP - A pipeline for RNA-Seq

Usually a combination of different tools might be desirable but, on the other side, time consuming since they depend on each others input/output and they are frequently focused in single purposes, hence often incompatible with one another [WMM+11].

iRAP [5] appears as an integrated solution that allows users to choose and apply their preferred tools for mapping reads, quantifying expression and testing for differential expression. [FPMB14]

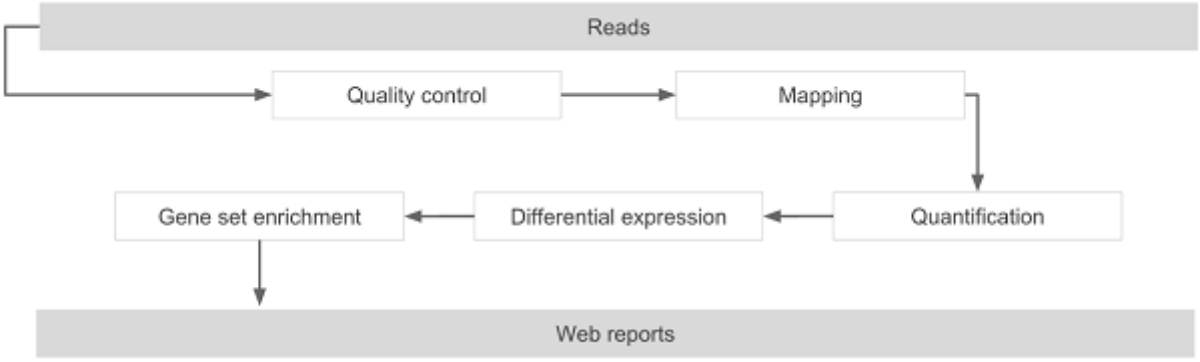

Figure 2.5: iRAP pipeline for RNA sequencing analysis.

The Figure 2.5 shows the main steps for this methodology that, by comparing with the previous workflow (Figure 2.4) has two more stages: Gene set enrichment and Web reports. Gene set enrichment is a means of analysing genomic data by studying the effects of groups of genes on a phenotype rather then individual gene analysis. [kn:10] Web reports facilitate the inspection of the results produced at different stages of the analysis by providing web pages with tables or plots of information [FPMB14].

iRAP is also capable of performing analysis starting at any stage of the pipeline by choosing the right command line options. Currently, it has the following five different stage options as described below:

setup/stage0 check all directories and data files are present and at the expected place;

stage1 quality filtering and reporting: check fastq file quality and report;

stage2 mapping: to the transcriptome or to the genome;

stage3 estimate gene expression;

stage4 estimate differential expression.

Furthermore, the expert can choose the specific tool for each phase of the system, some of which are described in detail in further sections. The current list of tools supported are available in Appendix A.

When using iRAP, the user has the possibility to speed up the analysis by using the `irap_lsf` command. This method accepts the same parameters as iRAP but splits the analysis into multiple jobs with the aim of reducing the time to analyze the data.

[5] iRAP repository: `https://code.google.com/p/irap/`

2.3 Standard File Formats used

For the implementation of the previous pipeline, some file formats are used to input or retrieve data. Below will be presented the most widely used ones along with some short description and detail when needed.

2.3.1 FASTA

FASTA [6] is the definition line and sequence character format used by NCBI [7]. It is a text-based format for representing either nucleotide (DNA or RNA) or amino acid (protein) sequences. FASTA is commonly used since it is easy to manipulate and parse using text-processing tools and scripting languages. Usually, in iRAP, a FASTA format is used to store the a reference genome.

2.3.2 FASTQ

FASTQ [8] provides a simple extension to the previous format: the ability to store a numeric quality score associated with each nucleotide in a sequence [CFG+10]. It is generally used as the first input with the raw data.

2.3.3 SAM and BAM

SAM [9] is a generic format for storing large nucleotide sequence alignments. It is easy to understand, easy to parse, and easy to generate and check for errors. Despite the presented qualities, SAM is slow to parse so a binary equivalent to it, BAM, has been developed to deal with this issue. BAM is used to intensive data processing and is useful in most product pipelines, while SAM has appropriate for interconversion with external applications and exploratory analysis. Like FASTQ, BAM/SAM formats are widely used as containers for raw sequence data.

2.3.4 GFF and GTF

GFF [10] is a format for describing genes and other features associated with DNA, RNA and Protein sequences. It was originally proposed as a protocol for the transfer of genomic feature information allowing people to develop and test them without having to maintain a complete feature-finding system. The second version also allows for feature sets to be defined over RNA and Protein sequences as well as genomic DNA.

On the other hand, GTF [11] stands for Gene transfer format and has been developed on top of GFF specification. The first eight GTF fields are the same as GFF. The group field has been

[6] FASTA format: http://genetics.bwh.harvard.edu/pph/FASTA.html
[7] NCBI: http://www.ncbi.nlm.nih.gov/
[8] FASTQ format: http://maq.sourceforge.net/fastq.shtml
[9] SAM and BAM file formats: http://samtools.sourceforge.net/SAMv1.pdf
[10] GFF format: http://www.sanger.ac.uk/resources/software/gff/spec.html
[11] GTF format: http://www.ensembl.org/info/website/upload/gff.html

expanded into a list of attributes. Each attribute consists of a type/value pair. Attributes must end in a semi-colon, and be separated from any following attribute by exactly one space.

2.4 Relevant Data Repositories

It is important to validate results and test along with the implementation of RNA-Seq pipeline. As such, real data are needed so that the quality and efficiency of this study is assessed. Some repositories are available online such as Gene Expression Omnibus (GEO [12]), Genomic Data Sharing (GDS [13]), ENSEMBL [14], UCSC Genome Browser [15] or the ENCODE Project [16],. The three last ones are highlighted in next section.

ENSEMBL

The Ensembl project provides information about many vertebrate and other eucaryotic species being started in 1999, some years before the draft human genome was completed. Until today, many genomes have been added over the years as well as the inclusion of comparative genomics, variation and regulatory data [HBB+02].

Between the included working teams of Ensembl, the Genebuild is the one who is responsible to create the gene sets both as developing and mantaining the BioMart [17] data mining tool.

It also supplies a REST server [18] as well as an FTP tool for browse and download data [CAB+15].

UCSC Genome Browser

The UCSC Genome Browser is a website that consists of an open-source tool to browse, analyze and query genomic data. Currently it grants access to some reference genomes, giving the possibility to retrieve specific data due to the search tool integrated.

Along with this tool it is also available an FTP server where data can be accessed and downloaded.

ENCODE Project

While ENSEMBL has information about many vertebrate and other species, ENCODE is a project to identify all functional elements in the human genome sequence as well as being able to create comprehensive, high quality catalogs of functional elements in the human genome using high–throughput technologies. Their mission is to enable the scientific and medical communities

[12] GEO: http://www.ncbi.nlm.nih.gov/geo/
[13] GDS: http://gds.nih.gov/02dr2.html
[14] ENSEMBL: http://www.ensembl.org/index.html
[15] UCSC Genome Browser: http://genome.ucsc.edu/
[16] ENCODE project: https://www.encodeproject.org/
[17] BioMart: http://biomart.org/
[18] REST server of Ensembl: http://rest.ensembl.org

to interpret the human genome sequence and apply it to understand human biology and improve health [The04].

Besides their motivation, an online repository is being build up, indexing gathered data and allowing users to query different data types regardless of location [Con11]. Some samples can be found at https://www.encodeproject.org/experiments/ENCSR274JRR/ for example.

2.5 Tools for RNA-seq and Differential Analysis

To perform the RNA-Seq pipeline some tools for each process are required. It is now presented some set of them on the following sections.

Bowtie

Bowtie [19] is a tool to align short reads to the human genome. To achieve speed and memory-efficiency, Bowtie aligns reads with the aid of an *index* of the reference genome [Lan10] and can take advantage of the use of simultaneous multiple processor cores [LTPS09]. For the human genome, Burrows-Wheeler is a commonly used indexing process that this tool extends with a novel quality-aware backtracking algorithm that permits mismatches.

TopHat

TopHat [20] is used as a junction mapping tool for RNA-Seq reads. Before TopHat, current mapping strategies could only localize reads to known exons in the genome. This tool is a read-mapping algorithm design to align reads to a reference genome without relying on know splice sites [TPS09].

Cufflinks

Cufflinks [21] assembles transcripts, estimates their abundances and tests for differential expression and regulation in RNA-Seq samples. Internally it is made of four different parts: Cufflinks that assembles the package, Cuffcompare that compares transcript assemblies to annotation, Cuffmerge that merges two or more transcript assemblies and Cuffdiff that takes the aligned reads and reports genes and transcripts that are differentially expressed [TRG+12a].

HTSeq

[19] Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
[20] TopHat: http://ccb.jhu.edu/software/tophat/index.shtml
[21] Cufflinks: http://cole-trapnell-lab.github.io/cufflinks/

HTSeq [22] is a python framework to facilitate the rapid development of scripts to processing and analysis of high-throughtput sequencing data.

FASTQC

FASTQC [23] is a quality control tool for high-throughput sequencing data. It receives as input BAM, SAM or FASTQ files, provides a quick overview on which areas there may be problems, summarizes graphs and tables for a fast assess of data and exports results to an HTML based report.

Samtools

Samtools [24] is a set of utilities to manipulate alignments in the BAM format (output format of TopHat, e.g.). It does sorting, merging and indexing, and allows to retrieve in any regions. The files can be either provided via standard input (stdin) of from a remote FTP or HTTP server.

deFuse

deFuse [25] is a software for gene fusion discovery from RNA-Seq data. It uses clusters of discordant paired alignments to inform a split read alignment analysis for finding fusion boundaries. As an output it produces a fully annotated output for each predicted fusion.

2.6 Relational and Non-relational Databases

Relational Databases had been introduced in 1970 by Edgar Codd at IBM Almaden Research Center [Cod70]. It has been used for decades and has brought in the concept of relations. With this, a relational database is usually represented by a structured model, a table, in which each row is a tuple (also considered as an object) and the columns are the attributes of the tuple. The database can be queried in order to retrieve useful data based on their attributes whether in common between tuples or not. To do so it is used a "structured query language", SQL [LM10]. The main set of operations is known as CRUD: Create, Read, Update and Delete. Relational databases are also know for their fixed schemas that force every tuple to have the exact same attributes.

A very common operation in SQL databases are joins. Joins are a way to associated data that share attributes or are somehow related. One example is imagining a table of people that have bank accounts associated and there is another table where the existing banks are described. A way to relate them is to combine these two tables in order to directly view which individual is using which bank. This is a very simple example that can scale very quickly and become a dangerous operation for the consumed resources to be used to compute the needed relationships.

[22] HTSeq: http://www-huber.embl.de/HTSeq/doc/overview.html
[23] FASTQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[24] Samtools: http://samtools.sourceforge.net/
[25] deFuse: https://bitbucket.org/dranew/defuse

In contrast with the previous set, a **Non-relational database** commonly doesn't have a fixed schema and avoid the previous complex operation. Within this class of databases there are different approaches to the used models such as: key-value stores, big table clones, document databases and graph databases. Each of them is used accordingly to the data set complexity.

MongoDB [26] is a non-relational database, schemaless and is scalable in cluster, performance and data. This way, it provides a way to store large data sets with the freedom of using different schemas according to data [27].

2.7 Tecnologies

When developing a web service, there are some basic resources and programming languages used. HTML [28] is a standard markup language for describing web documents. CSS [29] is used for defining how the HTML elements are to be displayed. Bootstrap [30] is a free front-end framework for faster and easier web development and contains HTML and CSS based design patterns. Javascript [31] serves as a method for programming the behavior of web pages. jQuery [32] is a library of Javascript that greatly specifies Javascript's usage.

Even though commonly used, PHP [33] is not the best suitable language for our thesis work. When compared to Python it lacks security and is not used on the Bioinformatics field. On the other hand, Perl has been greatly used on this field but it is currently being replaced by Python [34]. Furthermore, this last one has libraries like BioPython [35] for internal processing of samples and also has support to web through frameworks like Django [36] or BottlePy [37].

Even though Django is trending, it has no support for NoSQL relational databases yet. BottlePy, for instance, is a lightweight web framework, simple to use and provides the ability to use MongoDB.

Concerning the situation, the combination of Python and BottlePy is the most suitable since it has:

- Starting to replace Perl in bioinformatics processing;
- Support for NoSQL databases (either the framework and the programming language);
- Lightweight usage;

[26] MongoDB: http://www.mongodb.org/
[27] MongoDB scalability: http://www.mongodb.com/mongodb-scale
[28] HTML: http://www.w3schools.com/html/
[29] CSS: http://www.w3schools.com/css/default.asp
[30] Bootstrap: http://getbootstrap.com/
[31] Javascript: http://www.w3schools.com/js/
[32] jQuery: https://jquery.com/
[33] PHP: http://php.net/
[34] Python: https://www.python.org/
[35] BioPython: http://biopython.org/wiki/Main_Page
[36] Django framework: https://www.djangoproject.com/
[37] BottlePy: http://bottlepy.org/docs/dev/index.html

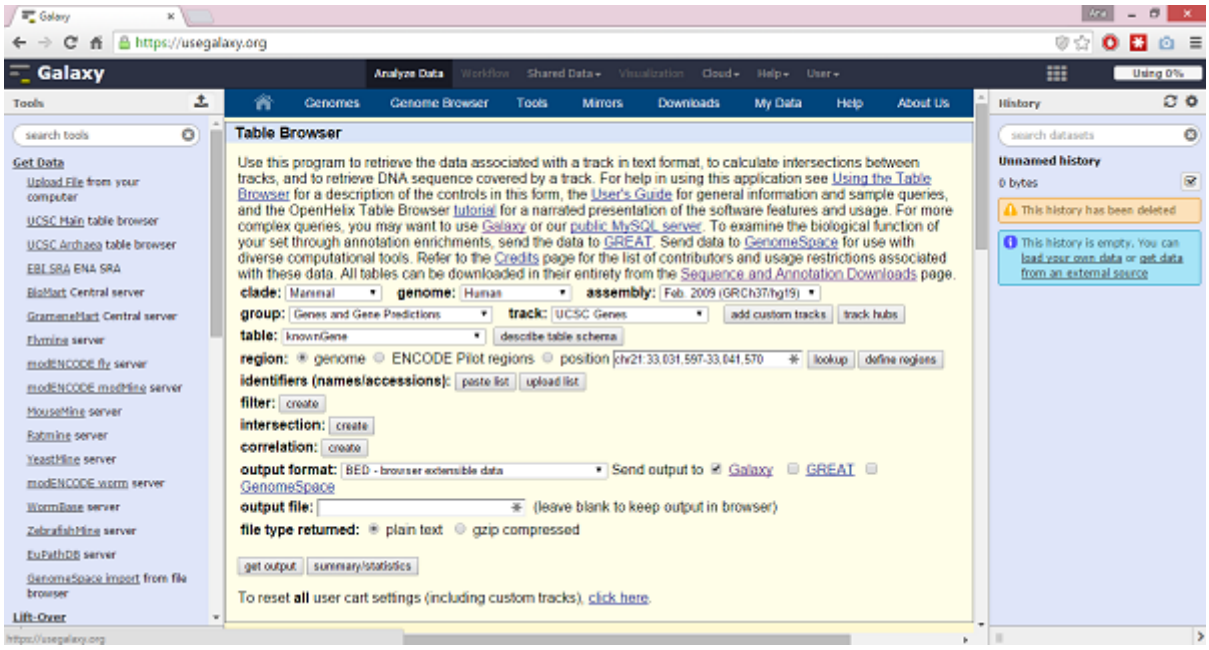

Figure 2.6: Print screen of Galaxy when trying to get data from UCSC.

.

For the purpose of sending the information between devices, the FTP server solution was analysed. It allows the creation of an administrator user that has the permissions for writing and reading but also for anonymous users that can only download data.

2.8 Related Work

Galaxy

Galaxy [38] is an open source, web-based platform for data intensive biomedical research.

It consists of an embedded website system that provides access to different databases such as Ensembl or UCSC Genome Browser. Although functional, the platform has not a consistent interface which requires the user to learn to use many systems within it. An example of this situation can be viewed in Figure 2.6. Furthermore it does not provide the ability to choose easily a stage pipeline, like iRAP, or to run a full one with personalized tools.

BioMart Central Portal

BioMart [39] is a community-driven project to provide unified access to distributed research data to facilitate the scientific discovery process.

[38] Galaxy: http://galaxyproject.org/

[39] BioMart Central Portal: http://central.biomart.org/sequence/?gui=Sequence_retrieval&mart=metaseq_mart_63_config

It currently has a sequence retrieval tools that provides searching by database and dataset. Nowadays, this tool offers Ensembl as the only available database but it adds up three other features like a sequences, filters and header information panels where the users can improve their search.

2.9 Chapter Summary

In this chapter we presented and discussed the technologies suitable for the thesis work. We have surveyed alternative technologies for gene expression computation and alternative splicing tools. The technologies are relevant to solve the Biological domain problem. We have also surveyed the tools for the development of websites and databases to store the data used when solving the biological problems. In the next chapter we describe work in detail.

Chapter 3

The Gemini framework

Despite the very large number of tools available for Molecular Biology problems, they are, most often difficult to use and require sophisticated computational resources. As part of the thesis work we have developed a framework that allows expert users to make RNA-Seq and alternative splicing analysis using a user friendly web interface and sophisticated computational resources. In this chapter we describe that framework that include the website, database resources for storing data and the computational settings for the analysis.

iRAP has been chosen to accomplish the automation of RNA sequencing analysis due to the simplicity of use and its wide number of tools available for its task. Since it is an integrated tool platform, its installation requires not only powerful computer resources but also a long list of basic equipment necessary to each of them. Considering these issues, it had not always been easy to setup the wanted work environment.

Despite the difficulties of this scenario, we have been able to complete the setup of this tool and make some control experiments with it.

In order to accomplish an iRAP execution, a configuration file must be provided (see Appendix B). Along with the experiment name, the name of species being analysed, the reference genome and read data files need to be specified and made available . Furthermore, the specific tools for each stage of the pipeline need also to be specified among other possible variables.

It is not easy and usually very time consuming for a Biologist to create the configuration file or to run the command-line interface that controls iRAP pipeline. Since this system is a combination of many different tools, there are a long list of available options that may be fulfilled. Some of them are mandatory while other ones are optional. After this setting is completed, it must be saved in a text file and made available to run iRAP.

Even though the usage of iRAP does not require extensive bioinformatics expertise, it does assume familiarity with UNIX command-line interface which would narrow down the target audience. In order to overcome this situation, we designed and developed a framework called Gemini

that includes a Graphical User Interface and makes RNA-Seq studies much more easy to perform. We have also included extra features for doing more analysis to RNA-Seq data.

3.1 Gemini

Due to the lack of intuitiveness on the process of generating the configuration file, it has been created a more user friendly one through a web form. This form asks for input of the main settings to call a very simple experiment with a commonly used bacteria: escherichia ecoli (e. coli). Another important feature is the possibility to fill the form asking what is the stage that the user wants the pipeline to start with.

Another issue was to find and easily download a reference genome for a certain species. In order to add completeness to Gemini and also to simplify the process of download, it has been implemented the functionality to download the reference genome (both fasta and annotation files). It is used the Ensembl API to obtain the required results. The jobs status feature lists the status of jobs sent by the user for execution.

Besides these features, one more have been added even though not completed. It is the *job results* where it is presented the results collected for a determined task.

3.2 Gemini Architecture

3.2.1 Physical Architecture

Gemini is divided into two different sets. The first one is where the web application is hosted as well as the database. Since dealing with genomic data often requires great computer performance, processing experiments is accomplished in an independent server that later responds with results.

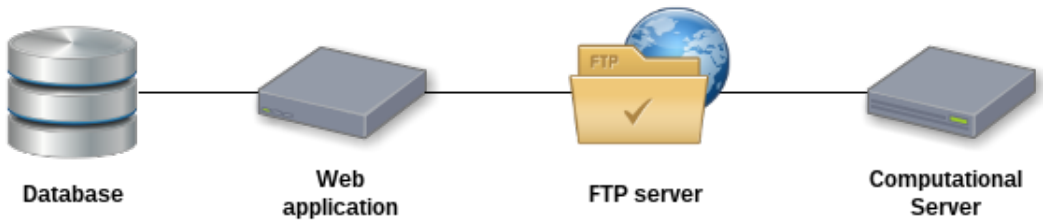

Figure 3.1: Physical architecture of Gemini.

Data scheme

MongoDB is NoSQL, document based, database. The information is organized in collections in which each element is stored as a JSON document. NoSQL databases are simpler to use when it

comes to have different information fields in the same document and, at the same time, they have better performance and scalability [SK11]. Currently the information is arranjed in two collections scheme as described on Table 3.1.

Name	Description
Accounts	Stores the users data.
Jobs	Stores the experiments data.

Table 3.1: Collections used to store data.

Accounts

An account's document stores an user's email, password and a unique identifier. A role has also been saved for future hierarchy creation in which, some levels will grant more permissions than others like changing accounts settings of other participants.

```
{
    "_id" : ObjectId("54f5a9b234f76511a82ec3b2"),
    "password" : "e10adc3949ba59abbe56e057f20f883e",
    "role" : "user",
    "email" : "acgg.gomes@gmail.com"
}
```

Jobs

Jobs are related to an user and have the name and description of the experiment. It also stores the reference genome's location as well as the reads' one. Likewise the date of creation is preserved. When a job is completed, the status is modified to "finished".

```
{
    "_id" : ObjectId("5544ecf134f7651d687bba94"),
    "name" : "ExpFinal",
    "exp_dir" : "./data/54f5a9b234f76511a82ec3b2/ExpFinal/reads",
    "user" : "54f5a9b234f76511a82ec3b2",
    "description" : "Final test",
    "genome_dir" : "./genomes/Escherichia coli",
    "status" : "started",
    "date" : ISODate("2015-05-02T15:27:45.249Z")
}
```

3.3 Data Management

In order to create a new job, it is required two sets of data, namely: *reference files* and *raw data files*. The raw data is the actual data being analysed while the reference genome serves as basis to the alignment or control.

The typical workflow consists on the expert providing the experience and species names as well as the raw data that was previously collected in the laboratory. Afterwards, through a step by step system, the user has the option to choose the reference genome to be used on the experiment. This last files are internally obtained from different FTP servers, such as Ensembl and ENCODE, stored locally and reused in further usages.

The provided information is then saved in the database and the files stored hierarchically. We use the following sequence of <*section name*> <*userid*> <*experience name*> <*data files*>. The <*section name*> can be *conf* for storing configuration files, *data* for saving the reads or *reference* for keeping the reference genome and annotation files. The <*userid*> is the ID of Accounts database document and the <experience name> is the same as *name* in Jobs database document. A directory tree is represented in Figure 3.2.

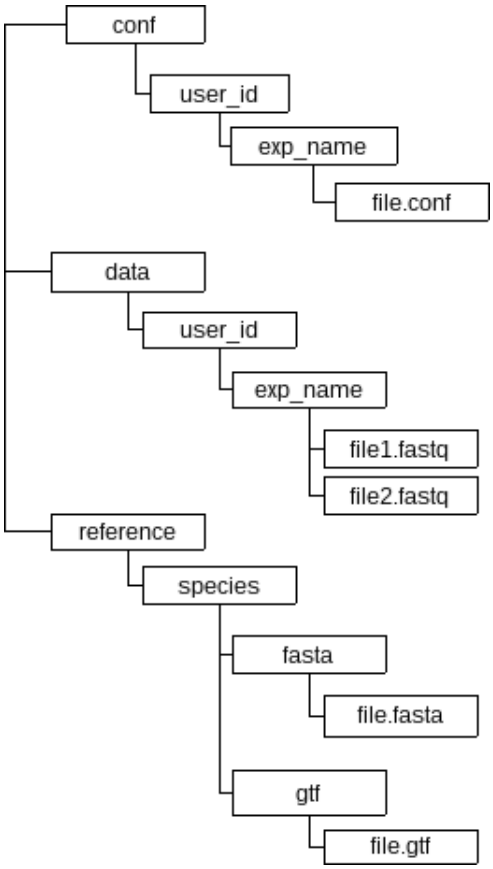

Figure 3.2: File system tree demonstrating the hierarchy of saved data files.

This information is then passed to the server via a local FTP server that has been created

for this purpose. We have followed an online tutorial [1] for the configuration. From this, we created one administrator that has permissions for writing and reading files and we also authorized anonymous login that may only retrieve files. This way, on the web application the admin user creates the reads, reference genomes, annotation and configuration files. On the computational server, an anonymous user authenticates for retrieving the files according to the filenames received via API.

The inside directory of FTP follows the hierarchy of <*userid*> <*experience name*> <*data files*>, where <*userid*> is the ID on the Accounts document and <*experience name*> is the name on Jobs document. When the server is prompted to run an experience, it downloads the required data from the FTP and runs iRAP with the necessary configuration options.

3.4 Web Interface

For the implementation of Gemini, python has been used for backend along with a simple, fast and lightweight WSGI micro web-framework, called BottlePy [2]. It has no dependencies and grants the main functionalities such as routing, templates, some class utilities and a built-in server.

The first screenshot (Figure 3.3) shows the home page of Gemini. There is top menu bar with two possible interactions: login and register. There is also a simple content box that describes the main features and status of development of the platform. Later on, a new menu appears with the possibility to logout on the *Welcome!* tab (as in Figure 3.4).

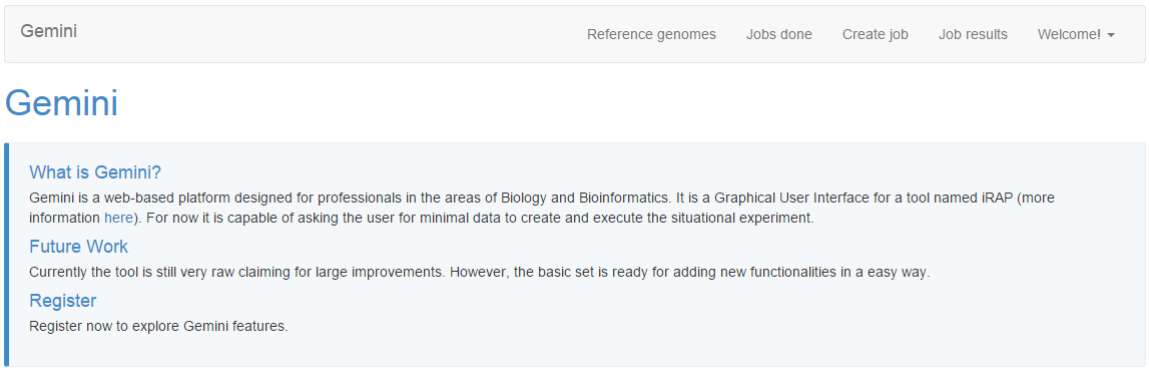

Figure 3.3: Home page.

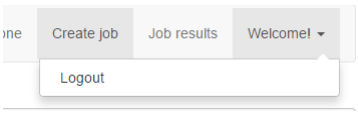

Figure 3.4: Logout menu.

[1] Tutorial available at: https://www.thomas-krenn.com/en/wiki/Setup_FTP_Server_under_Debian

[2] BottlePy: http://bottlepy.org/docs/dev/index.html

The main feature of the system is the *create job* (Figure 3.5) option that is easily accessible on top menu. When the user clicks on it, a form is displayed with the variables to introduce on the configuration file that iRAP needs. A description option has been added to provide the freedom to the researcher to include some extra information that he finds useful.

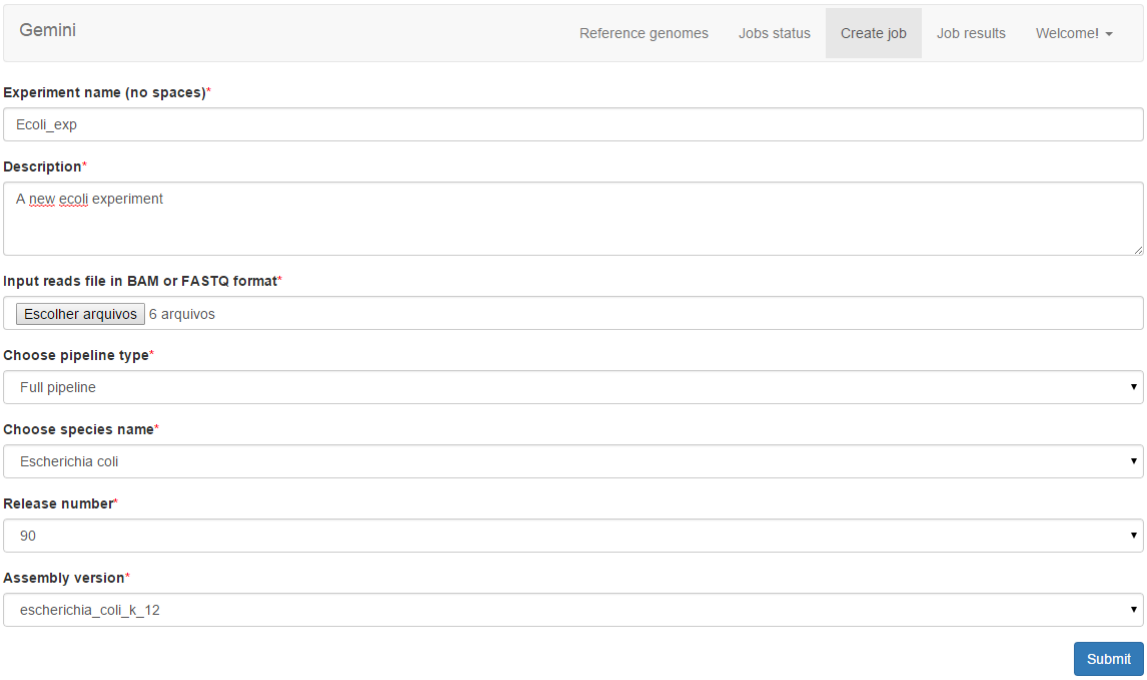

Figure 3.5: Create job form.

The same type of form was used for the reference genomes section (Figure 3.6). The user is prompted to choose, orderly, the species name, release number and version. Afterwards, a table of elements appears allowing the expert to download the files found.

Figure 3.6: Reference genomes component.

3.5 Use cases

Gemini was designed to have only one actor, the user that will accomplish the experiments. This actor has the possibility to register, login and logout of his account, create a job and the hability to download a reference genome of his choice. Even though not fully implemented, it has been considered the jobs done functionality.

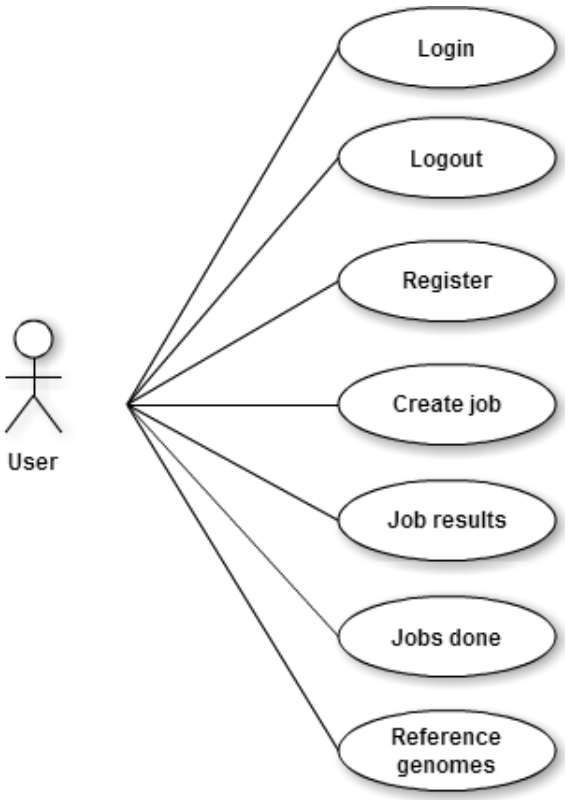

Figure 3.7: Use cases diagram.

3.6 Case studies

Some case studies with e. coli in order to test iRAP's installation and its capabilities. For both of the case studies, it has been collected six samples of Escherichia coli reads from European Nucleotide Archive [3] and the corresponding reference genome and annotation file from Ensembl.

3.6.1 Full pipeline

In order to call the full pipeline, iRAP has been tested with the directive `irap conf=ecoli_ex.conf`. Configuration file (*ecoli_ex.conf*) content is as described in Appendix C. In this task, the system received the reads and performed the quality control (QC), mapping, quantification and differential expression between the two groups of samples.

[3] ENA: http://www.ebi.ac.uk/ena

This test has been accomplished successfully.

3.6.2 Stage pipeline

The stage pipeline is a powerful functionality since it allows the user to choose a specific stage of the pipeline that he wants to execute for his samples.

Besides the fact that it is not easy to collect data samples, users often want to repeat or analyse just one step of the pipeline and take some conclusions of it. Furthermore, this type of pipeline is less time consuming.

A concrete usage example is IPATIMUP's data which contains the output information of the mapper TopHat and will be further analysed in detail in Chapter 4. Concerning this setup it would not be possible to call the full pipeline but instead it would require a command specific command option.

For validation purposes, the same bacteria from the previous analysis has been used for the stage pipeline case. After calling the setup command to check input, the directive `irap conf= ecoli_ex.conf quality_filtering_and_reporting` has been used to perform the *stage1* of iRAP. The platform was able to perform with success this task.

3.7 Chapter Summary

It has been used Python with BottlePy for the main implementation of the interface of Gemini. MongoDB has been chosen for dealing with information. A FTP server was mounted for the transitions of data between the client and the server. The platform allows the expert to call a full or stage pipeline analysis and provides the download of both fast and annotation files. The essential items are setup for the addition of further functionalities.

Chapter 4

Case study

In this chapter it is going to be presented the Case Study with data from samples of thyroid carcinoma along with their charactering and researching results.

After the development of an interface to the iRAP tool, the following step was to do a case study with some realistic DNA samples of thyroid cancer patients regarding the partnership with IPATIMUP's researchers.

The purpose of this case study was to do some research of potential differences between distinct types of the disease. In order to accomplish this, four samples were collected from Centro Hospitalar de São João (Porto). Cases 1 and 2 have been categorized as a widely invasive Follicular Thyroid Carcinoma (wFTC) and cases 3 and 4 as a minimally invasive Follicular Thyroid Carcinoma (mFTC).

4.1 Samples characterization

Patient from case 1 was female, with 82 years old, the tumour measuring 2 cm of size and presented a predominant follicular growth pattern and oncocytic features. Patient from case 2 was male, with 55 years old, the tumour measuring 5 cm of size and presented a predominant solid/trabecular growth pattern. For the group of mFTC, patient from case 3 was female, 56 years old, the tumour measuring 3.7 cm of size and with follicular growth pattern; patient from case 4 was female, 56 years old, the tumour measuring 4 cm of size and with follicular growth pattern.

4.2 Researching objectives

The aim of the study was to answer the following questions:

- Differentially expressed genes;
- Total isoforms found, either in percentange and in absolute;

- Average of alternative isoforms for each gene spliced;
- Potentially novel isoforms and how do they behave between samples/groups of samples;
- Compare results between versions of human genome assembly;
- Retrieve fusion genes;

4.3 Protocol of analysis

Collected sample reads have been aligned against reference genome GRCh37 with TopHat version 1.4.1. Posterior analysis followed Tuxedo Protocol as described in Trapnell et al. [TRG+12b]. An overview of the full protocol can be found in Figure 4.1.

For the implementation of this procedure, it were considered as two different groups the cases 1 and 2 (wFTC) against cases 3 and 4 (mFTC). The following step was to assemble transcripts for each sample with Cufflinks [1] and afterwards create a single merged transcriptome annotation with Cuffmerge [2]. Cuffdiff [3] was then used to find significant changes in transcript expression.

In order to find potential isoforms, each sample was ran against the reference genome with Cuffcompare [4].

4.4 Genome version GRCh37 versus GRCh38

Even though IPATIMUP's data have been assembled with reference genome version GRCh37, nowadays a more recent version (GRCh38) has been released in 2013. This new assembly is the first major revision of the human genome in more than four years [5].

The main differences between this two versions are:

Alternate sequences several human chromosomal regions have sufficient variability to prevent adequate representation by a single sequence. This new version provides alternate sequence for selected variant regions through the inclusion of alternate loci [6] scaffolds.

Centromere [7] representation Previous method of representing centromeric regions have been replaced by sequences from centromere models. The models provide the approximate repeat number and order for each centromere and are useful for read mapping and variation studies.

Mitochondrial genome The new mitochondrial reference sequence is the Revised Cambridge Reference Sequence with RefSeq accession number NC_012920.1. The previous one was RefSeq accession number NC_001907 which was not updated when GRCh37 assembly later transitioned to the new version.

[1] Cufflinks: http://cole-trapnell-lab.github.io/cufflinks/
[2] Cuffmerge: http://cole-trapnell-lab.github.io/cufflinks/cuffmerge/index.html
[3] Cuffdiff: http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html
[4] Cuffcompare: http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html
[5] As in UCSC mailing list at https://groups.google.com/a/soe.ucsc.edu/forum/#!topic/genome-announce/52Kv41YBXNY
[6] Specific location of a gene on a chromosome.

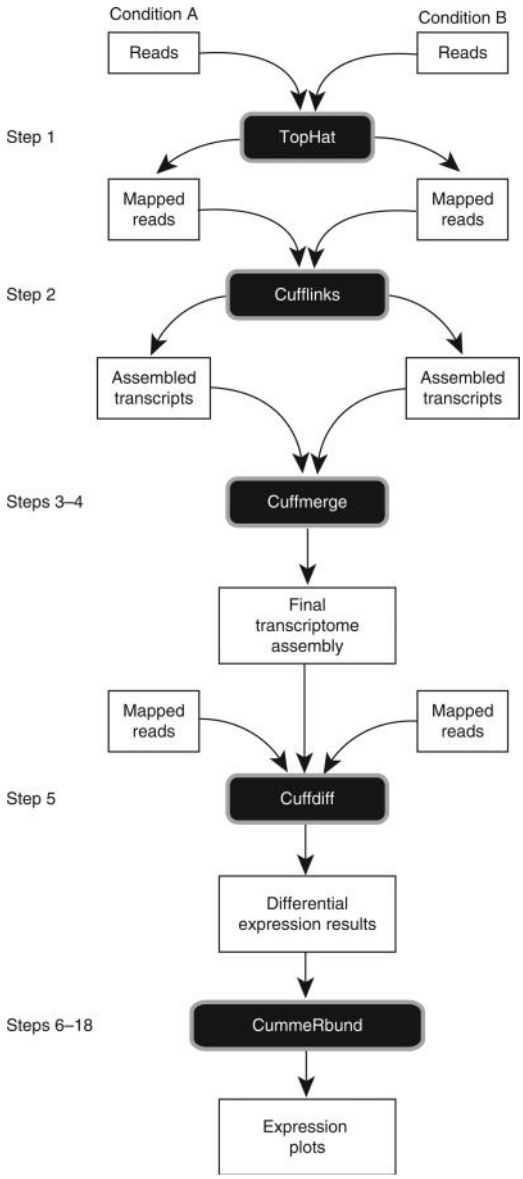

Figure 4.1: An overview of the Tuxedo Protocol with full steps: TopHat for alignment, Cufflinks package to find differential features between group samples and in the end CummeRbund to plot previous results.

Sequence updates Several erroneous bases and misassembled regions have been corrected and more than 100 gaps have been filled or reduced.

Analysis set The new assembly offers an "analysis set" that was created to accommodate next generation sequencing read alignment pipelines. Several GRCh38 regions have been eliminated from this set to improve read mapping.

4.5 Fusion genes

After we done some research on the technologies used for finding fusion genes in the reads we have not been able to perform the analysis to answer the last question of our researching objectives. DeFuse required as input the original reads which were not provided. This analysis would have been interesting for relating the ocurrence of thyroid and the presence of fusion genes as well as if there was any relevant difference between the samples of wFTC and mFTC.

4.6 Results

4.6.1 Case study with reference genome GRCh37

The following table (4.1) demonstrates some genes that were differentially expressed with significance between mFTC and wFTC. According to Celestino et al. [8] CRABP1, C1QL1 and LCN2 are biomarkers of thyroid cancer and predictors of extrathyroidal extension. It is commonly used the value of 0.05 has threshold above which the differential expression is significant. By inspecting out the table, we can find C1QL1 with a q-value of 0.000555 and CRABP1 with a q-value of 1.59E-11 which are in concordance with the previous article. Furthermore, it is verified the increased gene expression in wFTC when compared to mFTC.

Gene ID	Gene	Chr.	mFTC	wFTC	Fold change (log2)	Q value
ENSG00000167748	KLK1	19	0.18672	111.924	9.22742	0
ENSG00000104725	NEFL	8	0.0196645	27.6961	10.4599	0
ENSG00000174145	KIAA1239	4	0.0307388	14.4265	8.87444	0
ENSG00000182379	NXPH4	12	0.08307	19.5786	7.88074	0
ENSG00000131094	C1QL1	17	0.121453	2.38957	4.29828	0.000555
ENSG00000142677	IL22RAI	1	15.0905	0.384783	-5.29345	0
ENSG00000138615	CILP	15	4.23569	0.052726	-6.32793	2.41E-11
ENSG00000141431	ASXL3	18	1.8938	0.01408	-7.07148	1.24E-09
ENSG00000166426	CRABP1	15	74.8375	0.367899	-7.66831	1.59E-11
ENSG00000214145	LINC00887	3	19.797	0.119327	-7.37421	0.000459

Table 4.1: Genes differentially expressed between minimally invasive follicular thryoid carcinomas (mFTCs) and widely invasive follicular thyroid carcinomas (wFTCs).

In response to the previous researching objectives, the Table 4.2 has the statistical results for taks two and three. From the data analysis we could conclude that less than 1% of genes have isoforms but each one has an average of 3.89 alternatives.

[8] Ricardo Celestino, Torfin Nome, Ana Pestana, Andreas M. Hoff, Pedro A. Gonçalves, Catarina Eloy, Eva Sigstad, Ragnhild A. Lothe, Trine Bjøro, Manuel Sobrinho-Simões, Rofl I. Skotheim, Paula Soares, *CRAPBP1, C1QL1 and LCN2 are biomarkers of thyroid cancer and predictors of extrathyroidal extension*, subbmited for publication

Total genes	63896
Total significant isoforms	689
Total unique genes with isoforms	179
Ratio of: unique genes with isoforms/genes (%)	0.28
Average of alternatives per gene with isoforms	3.89

Table 4.2: Statistical results for researching objectives number 2 and 3.

4.6.2 Splicing analysis

Forward in the analysis, Cuffcompare [9] has been used to search for novel isoforms when compared to the reference genome. The method was to compare each sample to the reference genome and afterwards make an overlap of the results. The main objective was then to infer about the differences among wFTC and mFTC.

	Sample 1 (wFTC)	**Sample 2 (wFTC)**	**Sample 3 (mFTC)**	**Sample 4 (mFTC)**
Total genes	30440	36862	35539	37534
Potential novel isoforms	7406	8450	8781	8830

Table 4.3: Potential novel isoforms for each sample of IPATIMUP's data.

As demonstrated on Table 4.3 average there is a percentage of 24% of potentially novel isoforms found (at least one spliced junction is shared with a reference transcript). There is been made an overlap of the results in order to encounter differences between wFTC and mFTC. The Table 4.4 presents the counting of differences between cases. A brief analysis allows us to say that we can not make any relationship between wFTC and mFTC. On average, there was a total of 3722 changes among this two sets. The average within the sets are 3554 for wFTC and 3992 for mFTC. Given this results, there was not a significant divergence to allow us to relate the splicing and the occurrence of a more invasive Thyroid Carcinoma.

Sets	**Changes**
S1, S2	3554
S1, S3	3609
S1, S4	3612
S2, S3	3847
S2, S4	3820
S3, S4	3929

Table 4.4: Number of differences between samples. Syntax used: S1 - Sample 1, S2 - Sample 2, S3 - Sample 3 and S4 - Sample 4.

[9] Cuffcompare: http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html

4.6.3 Case study with reference genome GRCh38

In order to respond to the last research objective, it has been applied the method of analysis has before with a more recent human genome assembly. After using the same protocol with Cufflinks, Cuffmerge and Cuffdiff sequentially, we have come to some conclusions.

The produced results of differential expression analysis for the same genes were very different. Genes that were highly significant before, like C1QL1 and CRABP1, now had no relevance (q-value > 0.05). As mentioned before, we had only access to TopHat's output that used version GRCh37 for the alignment and mapping process, so we could not start from scratch the analysis.

This outcome demonstrates two topics in particular:

1. the introduction of novel annotations can make previous analysis outdated;
2. there is the need to save the original reads in order to update earlier studies.

The solution to this situation would have been to use the original reads to repeat the assembly and then follow the Tuxedo protocol to collected correct and updated differential expression data.

4.7 Chapter Summary

These section of development has been introduced in order to confer not only validation but also to do some researching studies on the subject. Four samples were collected and divided into two distinct groups. Tuxedo protocol was followed for responding to the research objectives. The case study for GRCh37 demonstrated the expression of C1QL1 and CRABP1 on wFTC when compared to mFTC as indicated in the referred article. On the other hand, the same protocol when applied with GRCh38 demonstrated the importance of saving the original reads and what is the influence of newer releases in previous researching results. Lastly, the splicing analysis results did not show any evidence of related isoforms count and the manifestation of the different stages of the disease.

Chapter 5

Conclusions and Future Work

The research of genomic-based cancers require highly complex data analysis. Such analysis require, most often, the use of a large number of complex computational resources, the gathering of a lot of information over the Internet and the use of powerful computational resources. Furthermore, the manipulation of the computational tools and resources, require the assistance of an informatics expert.

The thesis work is a contribution towards the solution of the above mention problem. We propose and developed a framework that enables an expert biologist to perform genomic-based complex analysis in a easy way and using the required powerful computational resources in a transparent way.

Using the developed framework we have used real data to address several research questions provided by IPATIMUP's researchers. Firstly, with the human genome GRCh37 we identified two of the biomarkers of thyroid cancer by differential gene expression analysis. Secondly, we found 689 significant isoforms, 179 unique, out of 63896 genes (ratio of 0.28%) with an average of 3.89 isoforms per gene. Thirdly, when trying to determine potential novel isoforms, we came at no conclusion of the differences of expression when comparing wFTC and mFTC. The last case was applying the Tuxedo protocol with the new GRCh38 on samples. The outcome of the experiment revealed the importance of saving the original reads and that newer releases become previous analysis outdated.

Future Work

This thesis work is just a starting point. As stated before the framework was designed so it can be extended with further modules to enable addressing other research questions.

The case studies also showed us that RNA-seq analysis is very time consuming. On the other hand, there are a lot of sub-tasks that can be run in parallel (like the task of aligning the reads).

We will study and compare different techniques and opportunities to make RNA-seq in parallel as much as possible.

References

[bNE] Scitable by Nature Education. Transcriptome. available at http://www.nature.com/scitable/definition/transcriptome-296, last access January 2015.

[CAB+15] Fiona Cunningham, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M.J. Searle, Giulietta Spudich, Stephen J. Trevanion, Andy Yates, Daniel R. Zerbino, and Paul Flicek. Ensembl 2015. *Nucleic Acids Research*, 43(D1):D662–D669, 2015.

[CFG+10] Peter J A Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71, April 2010.

[CH07] Geoffrey M. Cooper and Robert E. Hausman. *The Cell - A molecular approach*. American Society for Microbiology, 2007.

[Cod70] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, June 1970.

[Con11] The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9(4):e1001046, April 2011.

[FMB14] N. A. Fonseca, J. A. Marioni, and A. Brazma. RNA-seq gene profiling - a systematic empirical comparison. Technical report, May 2014.

[FPMB14] Nuno A. Fonseca, Robert Petryszak, John Marioni, and Alvis Brazma. irap - an integrated rna-seq analysis pipeline. *bioRxiv*, 2014.

[GCW+10] Qiang Gan, Iouri Chepelev, Gang Wei, Lama Tarayrah, Kairong Cui, Keji Zhao, and Xin Chen. Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. *Cell research*, 20(7):763–83, July 2010.

[GEN] GENIE. Dna, genes and chromossomes. University of Leicester, available at http://www2.le.ac.uk/departments/genetics/vgec/schoolscolleges/topics/dna-genes-chromosomes, last access January 2015.

[Gri] Malachi Griffith. Bioinformatics for Cancer Genomics (BiCG). available at http://bioinformatics.ca/workshops/2012/bioinformatics-cancer-genomics-bicg, last access January 2015.

[GTBK11] Angela Goncalves, Andrew Tikhonov, Alvis Brazma, and Misha Kapushesky. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics (Oxford, England)*, 27(6):867–9, March 2011.

[HBB+02] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.

[HBD+13] R. Hitzemann, D. Bottomly, P. Darakjian, N. Walter, O. Iancu, R. Searles, B. Wilmot, and S. McWeeney. Genes, behavior and next-generation RNA sequencing. *Genes Brain Behav.*, 12(1):1–12, Feb 2013.

[Ins] National Human Genome Research Institute. Transcriptome. available at http://www.genome.gov/13014330, last access January 2015.

[Joh03] Jason M. Johnson. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302:2141–2144, December 2003.

[kn:10] Gene set enrichment; a problem of pathways. *Briefings in functional genomics*, 9(5-6):385–90, December 2010.

[kn:11] Overview of available methods for diverse RNA-Seq data analyses. *Science China. Life sciences*, 54(12):1121–8, December 2011.

[Lan10] Ben Langmead. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 11:Unit 11.7, December 2010.

[LM10] Adam Lith and Jakob Mattsson. Investigating storage solutions for large data - a comparison of well performing and scalable data storage solutions for real time extraction and batch insertion of data. Master's thesis, 2010.

[LTPS09] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, January 2009.

[ORY10] Alicia Oshlack, Mark D Robinson, and Matthew D Young. From RNA-seq reads to differential expression results. *Genome biology*, 11(12):220, January 2010.

[SG15] Marc W Schmid and Ueli Grossniklaus. Rcount: simple and flexible RNA-Seq read counting. *Bioinformatics*, 31(3):436–437, February 2015.

REFERENCES

[SK11] Christof Strauch and Walter Kriha. NoSQL databases. *Lecture Notes, Stuttgart Media University*, February 2011.

[SOdMVN12] Mikhail Shugay, Iñigo Ortiz de Mendíbil, José L. Vizmanos, and Francisco J. Novo. Genomic hallmarks of genes involved in chromosomal translocations in hematological cancer. *PLoS Comput Biol*, 8(12):e1002797, 12 2012.

[The04] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–40, October 2004.

[TPS09] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11, May 2009.

[TRG+12a] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–78, March 2012.

[TRG+12b] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–78, March 2012.

[WMM+11] Ying Wang, Gaurang Mehta, Rajiv Mayani, Jingxi Lu, Tade Souaiaia, Yangho Chen, Andrew Clark, Hee Jae Yoon, Lin Wan, Oleg V Evgrafov, James A Knowles, Ewa Deelman, and Ting Chen. RseqFlow: workflows for RNA-Seq data analysis. *Bioinformatics (Oxford, England)*, 27(18):2598–600, September 2011.

[Wol13] Jochen B. W. Wolf. Principles of transcriptome analysis and gene expression quantification: an rna-seq tutorial. *Molecular Ecology Resources*, 13(4):559–572, 2013.

[ZW09] Mark Gerstein |& Michael Snyder Zhong Wang. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, January 2009.

REFERENCES

Appendix A

iRAP tools supported

A.1 Mappers

Mapper	Option name (mapper)
TopHat 1	tophat1
TopHat 2	tophat2
SMALT	smalt
GSNAP	gsnap
SOAPsplice	soapsplice
BWA	bwa1
BWA	bwa2
Bowtie 1	bowtie1
Bowtie2	bowtie2
GEM	gem
STAR	star
OSA	osa
MapSplice	mapsplice

A.2 Quantification

Quantifier	Option name (quant_method)
FluxCapacitor	flux_cap
HT-Seq (union mode)	htseq1
HT-Seq (intersection non-empty mode)	htseq2
Cufflinks 1 (try to find novel transcripts/-g)	cufflinks1
Cufflinks 2 (try to find novel transcripts/-g)	cufflinks2
Cufflinks 1 (do not assemble novel transcripts/-G)	cufflinks1_nd
Cufflinks 2 (do not assemble novel transcripts/-G)	cufflinks2_nd
NURD	nurd

A.3 Differential expression [DE]

Method	Option name (de_method)
DESeq	deseq
DESeq2	deseq2
EdgeR	edger
Cuffdiff 1	cuffdiff1
Cuffdiff 2	cuffdiff2

A.4 Gene set enrichment analysis [GSE]

Tool	Option name (gse_tool)
Piano	piano

Appendix B

iRAP model of configuration file

```
# This is a comment (lines that start with a # are ignored by iRAP)
# ============================================================================
# name that you want to give to the experiment (no spaces)
# All files produced by irap will be placed in a folder with the given name.
name=myexp


# ============================================================================
# name of the species
species=homo_sapiens

# ============================================================================
# Fasta file with the reference genome
reference=Homo_sapiens.GRCh37.66.dna.fa

# ============================================================================
# GTF file with the annotations
gtf_file=Homo_sapiens.GRCh37.66.gtf

# ============================================================================
# IRAP options (may be provided/overriden in the command line)

#######################################
# Mapper
#mapper=

#######################################
# Quantification method
#quant_method=

#######################################
# Dif. expression method
# Requires: contrasts to be defined (see below)
```

```
#de_method=

#####################################
# Gene set enrichment (GSE) analysis
# Requires: de_method to be defined
# Tool to use (none|piano)
# gse_tool=piano

# gse_method: (mean|median|sum|fisher|fisher-exact|stouffer|tailStrength|wilcoxon|
    reporter|page) see Piano vignete documentation for more details
#gse_method=fisher
#gse_pvalue=0.05
# Minimum number of genes in a gene set
#gse_minsize=3



#  TSV file with gene annotation
#  Format: "ID","Name","locus","source","lname","GO","GOterm","KEGG"
#  Description:
#  ID=gene_id (this is mandatory and should match the one given in the gtf file)
#  Name=gene name
#  locus=chr:start--end
#  source=biotype
#  lname=gene name
#  GO=go ids (multiple values can be separated by,)
#  GO=GO terms (multiple values can be separated by,)
#  KEGG=Kegg ids (multiple values can be separated by,)
#  If auto is defined then iRAP will *try* to generate the file
#  - this may take a considerable amount of time and will only work
#    for a reduced number of species
#  off - default value
#annot_tsv=

#####################################
# QC
# Check data (reads) quality (on|off|none)
# on  - reads are filtered out based on their quality
# off - the quality of the reads is assessed but no filtering is done
# none- no quality control is performed
#qual_filtering=on

# Trim all reads to the minimum read size after quality trimming -  Yes (y)| No (n)
# only applicable if qual_filtering is on
#trim_reads=y

# Minimum base quality accepted (def. 10)
#min_read_quality=10

```

```
82 # Contamination check (cont_index parameter).  Reads that likely
83 # originate from organisms other than the one under study can be
84 # discarded during pre-processment of the reads. This is done by
85 # aligning the reads to the genomes of organisms that might be a
86 # source of contamination and discard those that map with a high
87 # degree of fidelity. By default iRAP will check if the data is
88 # contaminated by e-coli. An example to create a contamination
89 # "database" is provided in the examples/ex_add2contaminationDB.sh
90 # script. The value of the parameter should be the file name prefix of
91 # the bowtie index files.
92
93 # Disable contamination check
94 #cont_index=no
95 # Default value
96 #cont_index=$(data_dir)/contamination/e_coli
97
98 ########################################
99 # Misc. options
100
101 # Number of threads that may be used by IRAP
102 #max_threads=1
103
104 # Exon level quantification ? Yes (y)| No (n)
105 #exon_quant=y
106
107 # Transcript level quantification? Yes (y)| No (n)
108 # transcript_quant=y
109 # ============================================================================
110 # full or relative path to the directory where all the data can be found.
111 data_dir=data
112
113 # the directory should be organized as follows (see directory data in IRAP toplevel
      directory)
114 #
115 # $data_dir
116 #$data_dir/
117 #          contamination
118 #                e_coli.1.ebwt
119 #                e_coli.2.ebwt
120 #                e_coli.3.ebwt
121 #                e_coli.4.ebwt
122 #                e_coli.README
123 #                e_coli.rev.1.ebwt
124 #                e_coli.rev.2.eb
125 #           raw_data
126 #                   $species
127 #                       .fastq
128 #                       .fastq
129 #                       ...
```

```
#           reference
#                $species
#                     $gtf_file
#                     $reference
#
# Notes:
#  1) $ denotes the value defined for the variable
#  2) Since version 0.5.0 the raw data (.fastq/.bam) files may be
# distributed across several sub-folders.


# ============================================================================
# Only necessary if you intend to perform Differential Expression analysis

# contrasts=contrast_def [contrast_def ...]
contrasts=purpleVsPink purpleVsGrey

# definition of each constrast
# contrast= group group [ group ...]
purlpleVsPink=Purple Pink
purlpleVsGrey=Purple Grey

# groups
# GroupName= Library_name [Library_name ...]
Purlple=myLib1 myLib2
Pink=myLib3
Grey=myLib4

# optional parameter: used in the report (HTML) generation.
#groups=Purple Pink Grey

# technical replicates
#technical.replicates="myLib1,myLib2;myLib3;mylib4"

# Note: names of groups, contrasts, and libraries should start with a letter and
    contain only alphanumeric characters and the character _.

# ============================================================================
# Data


# Information for each library
# LibName=Fastq file
# Note:
# 1. LibName should start with a letter and contain only alphanumeric characters
    and the character _. LibName should not contain in _1 or _2.
# 2. LibName should be different from the name of the fastq file, for instance
# f1=f1.fastq
# will produce an error.
```

```
177
178 # Single-end
179 myLib1=f1.fastq
180 # read size
181 myLib1_rs=75
182 # quality encoding (33 or 64)
183 myLib1_qual=33
184
185 # strand specific protocol?
186 #mylib1_strand=first
187 #mylib1_strand=second
188 # Default value is both (strands)
189 #mylib1_strand=both
190
191 # Have the file in a different sub-folder
192 # mylib1_dir=somesubfolder
193
194 # See SAM/BAM specification for more details about the following two parameters
195 # read group id (to be included in the BAM file) - this is not supported by all
      mappers
196 # myLib1_rgid=
197 # sam/bam header lines to include in the BAM file
198 # myLib1_shl="@CO\tThis is a comment\n@CO\tand another line..."
199
200 # LibName=Fastq file
201 myLib2=f2.fastq
202 # read size
203 myLib2_rs=75
204 # quality encoding (33 or 64)
205 myLib2_qual=33
206
207
208 # Paired-end
209 # LibName=Fastq files
210 myLib3=f3_1.fastq f3_2.fastq
211 # read size
212 myLib3_rs=50
213 # quality encoding (33 or 64)
214 myLib3_qual=33
215 # insert size
216 myLib3_ins=350
217 # standard deviation
218 myLib3_sd=60
219
220 # LibName=Fastq files
221 myLib4=f4_1.fastq f4_2.fastq
222 # read size
223 myLib4_rs=50
224 # quality encoding (33 or 64)
```

```
myLib4_qual=33
# insert size
myLib4_ins=350
# standard deviation
myLib4_sd=60




# list the names of your single-end (se) and paired (pe) libraries
se=myLib1 myLib2
pe=myLib3 myLib4
# No SE data
# se=
# No PE data
# pe=

#################################################################
#
# Passing/overriding parameters

# It is possible to pass parameters to the mappers and quantification
# methods but that should be done carefully since it may break the
# pipeline (if the location of the input and/or output files is
# changed).

# Overriding/changing the mappers' parameters:
# _map_options=options
# Example:
# tophat2_map_options=--min-intron-length 5 --no-coverage-search

# Overriding/changing the parameters of the quantification methods:
# _params=options
# Example
# htseq_params= -q
```

Appendix C

iRAP example configuration file

```
# minimal configuration file for DE
# experiment name
name=ecoli_ex
# species
species=ecoli_k12
# reference genome
reference=Escherichia_coli_str_k_12_substr_mg1655.GCA_000005845.1.19.dna.toplevel.
    fa.gz
# gtf file
gtf_file=Escherichia_coli_str_k_12_substr_mg1655.GCA_000005845.1.19.gtf.gz
# Enable filtering based on quality
qual_filtering=on
# Use a contamination data set to filter out reads
cont_index=no
# Toplevel directory with the data
# data_dir=

# some contrasts...
# GA=Group A
contrasts=GAvsGB GBvsGA
GAvsGB=GA GB
GBvsGA=GB GA
GA=FA FB FC
GB=FD FE FF

se=FA FB FC FD FE FF

FA=SRR933983.fastq.gz
FA_rs=50
FA_qual=33

FB=SRR933984.fastq.gz
FB_rs=50
```

```
FB_qual=33

FC=SRR933985.fastq.gz
FC_rs=50
FC_qual=33

FD=SRR933989.fastq.gz
FD_rs=50
FD_qual=33

FE=SRR933990.fastq.gz
FE_rs=50
FE_qual=33

FF=SRR933990.fastq.gz
FF_rs=50
FF_qual=33
EOF
```